Regresión Logística, entropía cruzada y divergencia de Kullback-Liebler

Pablo Musé

September 20, 2018

Consideremos una muestra etiquetada en particular $\{\mathbf{x}_n, \mathbf{t}_n\}$, del total de N muestras etiquetadas independientes. La idea es ajustar $\mathbf{w}$ para que la muestra $\mathbf{x}_n$ verifique:

$$P(\mathbf{x}_n \in C_1) = \sigma(\mathbf{w}^T\mathbf{x}_n) = y_n \geq 0.5 \qquad \text{si } t_n = 1,$$
$$P(\mathbf{x}_n \in C_2) = 1 - \sigma(\mathbf{w}^T\mathbf{x}_n) = 1 - y_n \geq 0.5 \quad \text{si } t_n = 0. \text{ (El caso 0.5 cae el la frontera de decisión).}$$

Cuanto más cercano está y_n a 1, más seguro es clasificar $\mathbf{x}_n$ en C_1; cuanto más cercano a 0, más seguro es clasificar $\mathbf{x}_n$ en C_2. Idealmente, querríamos que $y_n \to 1$ si $t_n = 1$, y $y_n \to 0$ si $t_n = 0$. Esto se puede resumir en una sola condición: queremos encontrar el $\mathbf{w}$ que haga que $y_n \to t_n$.

La etiqueta t_n es una variable determinística, pero podemos pensarla como una V.A. que, cualquiera sea el valor 1 o 0 que toma, lo hace con probabilidad 1. Dicho de otra manera, si $t_n = 1$ ($t_n = 0$), entonces sabemos con total certeza que $\mathbf{x}_n$ está en la clase C_1 (C_2). Haciendo abuso de notación:

$$\text{caso } t_n = 1 : \text{“}P(t_n \in C_1)\text{”} = P(t_n = 1) = 1 = t_n,$$
$$\text{caso } t_n = 0 : \text{“}P(t_n \in C_2)\text{”} = P(t_n = 0) = 1 = 1 - t_n.$$

De esta manera, buscamos el $\mathbf{w}$ que nos asegure que las densidades de probabilidad de las variables t_n y $\mathbf{x}_n$, que son respectivamente $p_{\mathbf{x}_n} = (t_n, 1-t_n)$ y $q_{\mathbf{x}_n} = (y_n, 1-y_n)$, sean los más parecidas posibles.

Una medida de cuán distintas son dos densidades es la divergencia de Kullback-Leibler. De forma general, si tenemos dos densidades $p = (p_1, p_2, ..., p_K)$ y $q = (q_1, q_2, ..., q_K)$, esta divergencia se define como

$$KL(p||q) = \sum_{k=1}^{K} p_k \log\left(\frac{p_k}{q_k}\right).$$

Se puede ver que $KL(p||q) \geq 0$, con la igualdad si y solo si $p = q$. La entropía cruzada entre dos V.A.s (o sus dos distribuciones), $H(p, q)$, se define como

$$H(p, q) = -\sum_{k=1}^{K} p_k \log(q_k).$$

Es fácil ver que

$$KL(p||q) = H(p, q) - H(p).$$

$H(p)$ es la entropía de p, pero esto es un detalle que podemos obviar a los efectos de entender la regresión logística; basta con ver que la densidad q que minimiza $KL(p||q)$ es la mismo que minimiza $H(p,q)$.

Volviendo ahora a la regresión logística (2 clases). Queremos que $p_{\mathbf{x}_n} = (t_n, 1-t_n)$ y $q_{\mathbf{x}_n} = (y_n, 1-yn)$, sean los más parecidas posibles, entonces por lo anterior queremos encontrar un $\mathbf{w}$ que minimice $KL(p||q)$, o en su defecto $H(p,q)$. Definimos entonces el costo siguiente para la muestra $\{\mathbf{x}_n, tn\}$, relativo a $\mathbf{w}$:

$$E_n(w) := H(p_{x_n}, q_{x_n}) = t_n logy_n + (1-t_n)log(1-y_n).$$

Cuando tenemos N muestras, minimizamos la entropía cruzada promedio sobre todas las muestras, o simplemente su suma (que es la log-verosimilitud de una distribución binomial, como vimos en clase):

$$E(w) := \sum_{n=1}^{N} E_n(w) = \sum_{n=1}^{N} t_n logy_n + (1-t_n)log(1-y_n).$$

En suma, el problema de regresión logística para dos clases con N muestras etiquetadas consiste en encontrar el $\mathbf{w}^*$ que clasifique con menor error posible el total de las muestras, en el sentido de la entropía cruzada promedio (o total):

$$w^* = \arg\min E(w).$$

Como vimos en clase, este costo es estrictamente convexo por lo que admite un único mínimo, y C^2 lo cual nos permite encontrarlo mediante un método de Newton (que en este caso resulta en el Iterative Reweighted Least Squares).